UNIVERSITY OF MORATUWA

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

B.Sc. Engineering
2009 Intake Semester 7 Examination

CS4342 ADVANCED COMPUTER ARCHITECTURE

Time allowed: 2 Hours September 2013

ADDITIONAL MATERIAL: *None*

INSTRUCTIONS TO CANDIDATES:

1. This paper consists of **5** questions in **6** pages.
2. Answer any FOUR (4) questions.
3. Start answering each of the main questions on a new page.
4. The maximum attainable mark for each question is given in brackets.
5. This examination accounts for 60% of the module assessment.
6. This is a closed book examination.

   NB: It is an offence to be in possession of unauthorised material during the examination.
7. Only calculators approved by the Faculty of Engineering are permitted.
8. Assume reasonable values for any data not given in or with the examination paper. Clearly state such assumptions made on the script.
9. In case of any doubt as to the interpretation of the wording of a question, make suitable assumptions and clearly state them on the script.
10. This paper should be answered only in English.

Answer any four (4) questions.

Question 1 (25 marks)

Suppose just after graduation you were hired by Apple. Your team is responsible for designing the next processor for the 8th generation iPhone to be released in 2016. iPhone 8 is to be the thinnest smart phone ever with high quality graphics, fast app performance, and extended battery life. As a preparation for the first project meeting, each team member is asked to think about how to implement various components of the processor.

(i) Briefly explain what type of an approach/technology you would recommend to each of the following items. Provide at least 2 justifications for each of your selections.

a) Number of processing elements (e.g., single core, multi core, or many core). [3]

b) Parallel execution of instructions (e.g., pipelines, vectors, multimedia extensions, many SIMD lanes, etc.). [4]

c) Memory hierarchy (e.g., number of levels, cache size, cache associatively, etc.). [6]

(ii) After several performance tests it has been identified that the I/O sub-system on the 7th generation iPhone was the primary performance bottleneck, where the processor waits for I/O 60% of the time. Therefore, one of your team members suggested that by speeding up the I/O system alone it would be possible to gain an overall speedup of 2×.

a) How much speed up in the I/O sub-system will be required to achieve the desired overall speed up?

You may use the Amdahl's law given below for the calculation:

$$Speedup_{Overall} = \frac{1}{(1 - Fraction_{Enhanced}) + \frac{Fraction_{Enhanced}}{Speedup_{Enhanced}}}$$ [3]

b) Do you agree with your colleague's idea of gaining such a speedup by improving only the I/O sub-system? Briefly explain. [3]

(iii) Social responsibility is a key aspect for a reputed brand like Apple. Briefly describe what would be your suggestion to build the greenest iPhone ever. Your suggestions should consider full life-cycle analysis based on green computing concepts. [6]

Question 2 (25 marks)

(i) What are the pros and cons of compile time loop unrolling? [4]

(ii) Consider the following Assembly code. The value of `R2` is predefined.

```
Loop:
      LD     F1,0(R1)
      LD     F2,100(R1)
      DADD   F3,F1,F2
      SD     F3,200(R1)
      DADDUI R1,R1,#-4
      BNZ    R1,R2,Loop
```

a) Identify all data dependencies in the above code. [3]

b) Calculate the number of clock cycles required by an iteration of the above loop. Use the following table for latencies of Floating Point (FP) operations in MIPS. [3]

Instruction Producing Result	Instruction Using Result	Latency in Clock Cycles
FP ALU operation	Another FP ALU operation	3
FP ALU operation	Store double	2
Load double	FP ALU operation	1
Load double	Load or Store double	0

c) Improve the above code to reduce the number of clock cycles. [3]

d) Recalculate the clock cycles required by an iteration of the new loop. [2]

e) Unroll the loop to further reduce the number of clock cycles. [8]

f) How much speedup can we gain per iteration compared to the original code? [2]

Question 3 (25 marks)

(i) Following write up is extracted from "Evaluating Intel's Many Integrated Core Architecture for Climate Science" by Theron Voran, Jose Garcia, and Henry Tufo.

Knights Ferry (KNF) is the first generation of Intel's Many Integrated Core Architecture. KNF is implemented on an ×16 PCIe 2.0 card plugged into a Xeon host system. The KNF card has up to 32 cores clocked up to 1.2 GHz, supporting 4 hardware threads per core and a short in-order pipeline. Each core has a 512-bit SIMD vector processing unit.

Each core is provided a 32KB L1 data cache and a 32KB L1 instruction cache, as well as a 256KB L2 cache. The L2 caches for each core are interconnected via a bidirectional ring bus, creating an 8MB globally-shared, coherent cache. Additionally the KNF cores share 1 or 2GB of GDDR5 main memory.

Based on the above description answer the following questions.

a) Altogether how many hardware-level threads are in KNF? [2]

b) Briefly explain what is mean by "short in-order pipeline" [2]

c) Briefly explain what is mean by "512-bit SIMD vector processing unit"? [2]

d) Briefly explain what is mean by "globally-shared, coherent cache"? [2]

e) Draw the cache/memory hierarchy while labelling necessary components and capacities. No need to draw all the cores. [5]

(ii) NVIDIA Tesla 2070 GPU has 32 SIMD lanes per processor. There are 14 SIMD processors per card. It allows running at most 512 threads per SIMD processor (block) at a given time.

a) Altogether how many cores are in the GPU card? [2]

b) Suppose we want to run 5,000 threads on this system. Explain how such a large number of concurrent threads can run on the Tesla 2070. [4]

(iii) Branch prediction can be performed at compile time, run time, or both.

a) Draw a 2-bit Branch Predictor. [3]

b) Is run time branch prediction is desirable for each SIMD lane? Explain. [3]

Question 4 (25 marks)

(i) Compare and contrast (i.e., similarities and differences) Static Scheduling and Dynamic Scheduling. [4]

(ii) Using suitable examples briefly describe 4 cache optimization techniques. [8]

(iii) Mobile Processors are now being used in Data Centers and Warehouse-Scale Computers. For example, ARM recently released several 64-bit processors targeting the server market.

Briefly describe what are the advantages and disadvantages of using Mobile Processors in such systems. [4]

(iv) Consider a Warehouse-Scale Computer (WSC) with 100,000 nodes. Mean Time To Failure (MTTF) of a node is 4 years.

a) What is the expected number of failures per day? [3]

b) Hardware and software failures are a norm rather than an exception in WSCs. Then how do WSCs achieve dependability/reliability? [2]

(v) Figure Q4 shows the power consumption with different workload according to the SPECpower benchmark.

Suppose an application is deployed on 3 servers each with 20% of the workload (from the maximum workload). Can we save actual power by consolidating the all 3 servers to a single server? Explain. Assume workloads are independent. [4]

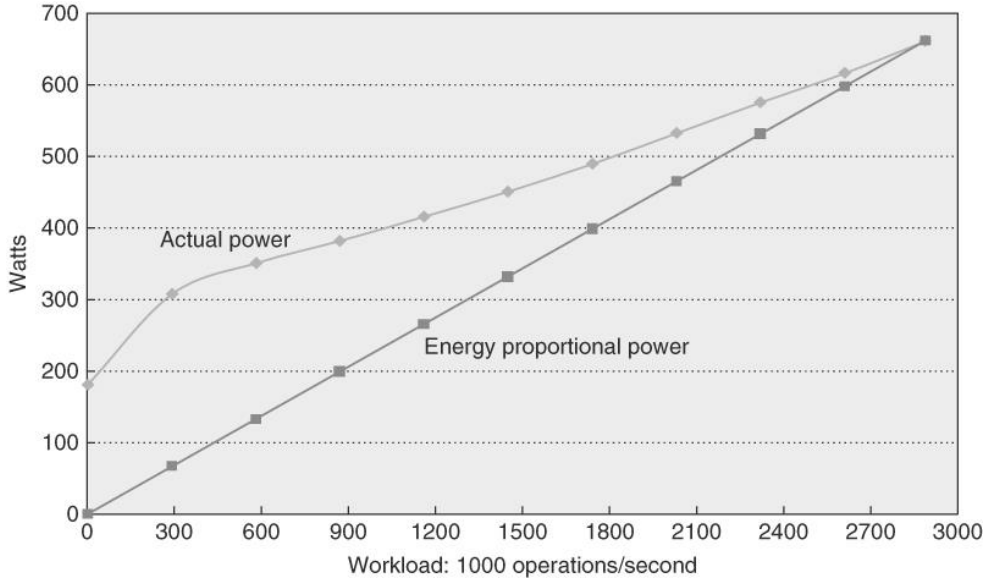

Figure Q4 – Workload vs. power consumption. Source – "Computer Architecture, A Quantitative Approach" by John L. Hennessy and David A. Patterson.

Question 5 (25 marks)

(i) List an advantage and a disadvantage of Direct mapped cache? [2]

(ii) On a given system 50% of the instructions are load and store. Miss penalty is 25 clock cycles and miss rate is 2%. If *n* instructions to be loaded and executed, how many clock cycles will be wasted due to memory stalls? Suppose Clocks Per Instruction (CPI) is 1. [3]

(iii) Answer any 4 of the following questions. These are based on the papers discussed in the class.

a) What the inclusion property is? Why is it important in cache coherence? Explain using a suitable example(s). [5]

b) How can hybrid cache architectures with disparate memory technologies provide better caching performance than today's caches? [5]

c) Today's GPUs have a very limited cache coherency support. How can we use Temporal Coherence to address the cache coherence needs of GPUs? [5]

d) What is Distant Instruction Level Parallelism (ILP)? How can it be exploited to gain better processor performance? [5]

e) While co-location of virtual machines on a warehouse scale computer provides many benefits, it also affects the Quality of Service (QoS) experienced by an application. Briefly explain how carefully designed stress tests and predictions can better schedule virtual machines while preserving the QoS? [5]

f) What are the challenges that we have to address while optimizing the energy consumption of a computer system? [5]

---------- END OF THE PAPER ----------